A Data Mining Based Approach for Determining the Potential Fishing Zones

Devi Fitrianah, Hisyam Fahmi, Achmad Nizar Hidayanto, and Aniati Murni Arymurthy

***Abstract*—The aim of this paper is to analyze the determination of the potential fishing zones based on data mining approach. The algorithm utilized in this study is AGRID+, a grid density based clustering for high dimensional data. The case study area is in eastern Indian Ocean located at 16.56 - 2 S and 100.49 – 140 E. The algorithm is implemented in 7 phases, partitioning, computing distance threshold, calculating densities, compensating densities, calculating density threshold, clustering and removing noise. The clustering result is evaluated by the Silhouette index. The results of the study show that the best cluster formed at daily aggregate temporal with number of cell (*m*) = 14 and the number of cluster formed was 50 clusters. The constant execution time is in line with the increasing the value of *m*. From three different temporal aggregate, the daily aggregate is running relatively constant for various *m* value. To determine the potential fishing zones for different temporal aggregate can be achieved by applying the thresholding technique to the cluster result. Utilizing the data mining approach yielded a prominent 22 daily clusters identified as potential fishing zone.**

***Index Terms*—AGRID+, data mining, clustering, potential fishing zone, spatio-temporal.**

I. Introduction

Indonesia is a maritime country with 5.8 million km square of ocean area and 81,000 km of long beach.Having the great potential in marine resources, especially in fishing, Indonesia is becoming the third largest fishery producer country based on the world statistics data in 2007. Tuna fishing is becoming the third largest national commodity in Indonesia. To further improve the tuna commodity, many approaches and techniques are developed, including the utilization of technology in determining the potential fishing zones. The most applicable technique is utilizing the geographic information system [1]. In line with geographic information system, Sadly *et al.*, utilizing the knowledge based expert system with some rules generated from characteristics to determine the potential fishing zone [2]. Other study is by creating a simple prediction map based on remote sensing data and fishery data [3]. The latest study on determining the potential fishing zone utilizing a clustering technique to set a rough map of potential fishing zone based on fishing data [4].

In this study, we need a clustering algorithm that can treat spatio-temporal data used to determine the potential fishing zone by grouping the fishing areas based on location and time. The grid-density based clustering method is used to cluster fishing area. This method is called AGRID+ proposed by Zhao *et al*. [5]. Referring to the clustering result, we can determine whether the area is potential area or not based on the criteria from expert.

The contribution of our paper is to propose an alternative approach to determine the potential fishing zone utilizing data mining technique.

Manuscript received August 19, 2014; revised October 22, 2014. This work was supported in part by the Universitas Indonesia Research Grant 2014 (BOPTN UI).

The authors are with the Faculty of Computer Science, Universitas Indonesia, Depok 16424 Indonesia. Devi Fitrianah is also with the Faculty of Computer Science, Universitas Mercu Buana, Jakarta 11650 Indonesia (e-mail: fitrianah.devi@gmail.com, hisyam@cs.ui.ac.id, nizar@cs.ui.ac.id, aniati@cs.ui.ac.id).

II. Literature Review

A. Spatio-Temporal Data Mining

One of the approaches in spatio-temporal data mining is a spatio-temporal clustering where the process is to analyze the data object without knowing the label of its class. Spatio temporal clustering itsef is widely used in many fields, e.g. in medical application areas, security, environment, biology, pathology, health, fisheries, and others [6].

This spatio-temporal clustering method can be implemented either by conducting a single step clustering in the 3-dimensional data or do the clustering in two phases, i.e. get spatial information first and then the temporal information. Each type of clustering give a slightly different results, because each type give emphasis on certain dimensions, whether in spatial dimension or in temporal dimension [7].

B. AGRID+

AGRID+ algorithm proposed by Zhao *et al.* [5] is effective for high-dimensional data clustering. AGRID+ is a grid-density based clustering which is improved from AGRID (Advanced Grid-based Iso-Density line clustering) algorithm [8]. It takes object as the smallest element in the clustering process and the addition of *i*th-order neighbor concept that considering low-order neighbor only to decrease the complexity of computation process. AGRID+ algorithm use the idea of density compensation to make up the accuracy loss caused by ignoring high-order neighbor. Density compensation is the product of its original density and ratio of the volume of the neighborhood to that of the considered part of neighborhood [5].

AGRID+ algorithm is consists of seven steps, partitioning, computing distance threshold, calculating densities, compensating densities, calculating density threshold, clustering and removing noise.

1) Partitioning. The whole data is partitioned into cells according to the number of cells in each dimension (*m*). Every non-empty objects are inserted into the cells with corresponding coordinates. The coordinates are computed based on the interval in each dimension (*L*).

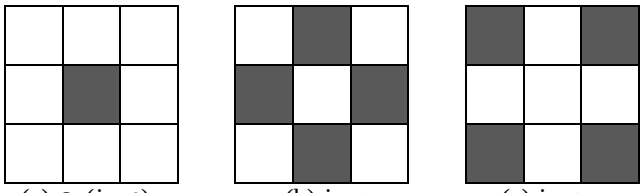
2) Computing distance threshold. The distance threshold (*r*) is a criteria to define whether two objects are close enough or not. It is computed according to the interval lengths of every dimensions using Equation (2).
3) Calculating densities. The density for each object is counted according to the number of objects both in its neighborhood and in its neighboring cells using *i*th-order neighbor concept.
4) Compensating densities. The density compensation is the product of its original density and ratio of the volume of the neighborhood to that of the considered part of neighborhood.
5) Calculating density threshold. The density threshold (*DT*) is the ratio between compensated densities and θ coefficient as shown in Equation (3). It is a criteria to define the minimum number of objects that should exist in a cluster.
6) Clustering. Firstly, each object that has a density greater than *DT* is labeled as a cluster. Then, for each object are checked with the other objects in neighboring cells. If the other object in the neighboring cells has a greater density than *DT* and its distance less than *r*, then that two clusters that contain that objects are merged into one cluster.
7) Removing noise. From those clusters obtained, clusters with the average density less than *DT* considered as a noise and will be removed.

This study implement the AGRID+ algorithm for clustering fish catch data which is the spatio-temporal data that has the information of fish catch coordinates as spatial dimension and fishing time as temporal dimension.

III. Methodology

A. Data

Data used in this study is spatio-temporal data about daily fish catch around the Indian Ocean between the years 2000 until 2004. The geographic coordinates for this location area ranging from latitude 16.56 – 2 S and longitude 100.49 – 140 E. This data obtained from fish catch data by the shipping company PT. Perikanan Nusantara Indonesia.



(a) C_a (i = 0) (b) i = 1 (c) i = 2

Fig. 1. The *i*th-order neighbor of C_a.

B. AGRID+ Clustering

In the AGRID+ algorithm, firstly each dimension is divided into many intervals and data is partitioned into hyper-rectangular cells, the 3d-rectangular cells used for fish catch spatio-temporal data. The interval value is calculated based on the number of 3d-rectangular cells to be formed. Interval values for each dimension will be different according with the number of rectangular in its dimension. The length of cell's interval in each dimension obtained from the dimension's range divided by the number of cells in that dimension. The computation of interval length can be formulated in Equation (1).

$$L_d = \frac{\max(data_d) - \min(data_d)}{m_d} \quad (1)$$

where, L_d is the interval length for each cells at *d*th-dimension, $data_d$ is the feature of the data at *d*th-dimension, and m_d is the number of cells at *d*th-dimension. We have to do features normalization before the clustering process due to the difference in scale between spatial features and temporal feature.

After the length of intervals for each dimension, then the partition is done by assigning each object to the cell according to the value of its features. Next step, compute the distance between an object a in a cell and the objects in its neighboring cells, and its density is the count of objects that are close to object *a*. Determination of the neighbor is based on the *i*th-order neighbor. Zhao *et al.* [5] defined the *i*th-order neighbor as a cell in *D*-dimensional space which shares $(D - q)$ dimensional facet with cell C_a, where cell C_a is a cell in the *D*-dimensional space that contain an object *a*, and q is an integer between 0 and *D*. The 0th-order neighbors of C_a is C_a itself. Examples of *i*th-order neighbors in a 2D space are shown in Fig. 1.

To determine the neighborhood of the object required an *r* parameter, the radius of neighborhood. Choosing the value of *r* is not so easy, because if *r* is large enough, this algorithm will become like a grid-based clustering. On the other hand, if *r* is too small, this algorithm will become like a density-based clustering. So, to determine this parameter, Zhao *et al.* said that the value of *r* must be less than *L*/2, where *L* is the minimum interval length in all dimensions [5]. The value of *r* can be calculated using Equation (2).

$$r = \lambda \times \frac{\min(L)}{2} + (1 - \lambda) \times \frac{\max(L)}{2} - \varepsilon \quad (2)$$

where, λ is a weight coefficient ($0 < \lambda < 1$), *L*is the interval length for all dimensions, and$\varepsilon$is a very small number, so that we achieved the value of $r < L/2$.

The clustering result also determined by the value of density threshold *DT* that can be computed with Equation (3).

$$DT = \frac{mean(Density_i)}{\theta} \quad (3)$$

where, θ is the coefficient that can be tuned to get the cluster results at different level. A small value of θ will lead to a big *DT* and vice versa. The appropriate cluster level can be decided by the needs of user itself. In this experiment, we use DT = 5 or equivalent with set θ = 1, based on the criteria to determining the potential fishing zone.

Continue to the clustering process, each object that has a density greater than *DT* is labeled as a cluster. Then, every pair of objects which are in the neighborhood of each other is checked. If that pair of objects are close enough (distance $\leq r$) and have a density that meets the criteria as a cluster (density $\geq DT$), then that two clusters that contain that pair

of objects are merged into one cluster.

C. Cluster Evaluation

Silhouette index is used to evaluate the clustering results. The values of silhouette index for each object indicate the representation of how well each object lies within its cluster. It was first described by Peter J. Rousseeuw [9]. Silhouette index is computed according to the similarity between an object and the other objects of the cluster it belongs to compared with the similarity between an object with the other objects of each of the other clusters. Equation (4) is the formula to compute silhouette index.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

which is called the silhouette width of the object, where $a(i)$ is the mean distance of *i*th-object to the other objects of the clusters it belongs to, and $b(i)$ is the smallest value of the mean distance of *i*th-object to the objects in other clusters.

The value of silhouette index is between -1 and 1. A value near 1 indicates that the object is affected to the right cluster. Otherwise a value near -1 indicates that the point should be affected to another cluster.

D. Potential Fishing Zone

Based on the fish catch area from clustering results, the determination of potential fishing zone (PFZ) can be done by computing the mean of fish catches for each area. Then, compare each of that means with the threshold. If it is greater the threshold, that area can be categorized as potential area. The threshold is from the expert opinion, i.e. the shipping company PT. Perikanan Nusantara Indonesia, which states that a catch point is categorized as potential if its number of fish catches is equal or greater than 5 in a single trip.

The thresholding technique to determine the PFZ computed in Equation (5).

$$PFZ = \begin{cases} 1, & \frac{\Sigma Catch_{ij}}{N_i} \geq th \\ 0, & otherwise \end{cases} \quad (5)$$

where $\Sigma Catch_{ij}$ is the total number of fish catches at *i*th-cluster, N_i is the number of catch points at *i*th-cluster, and th is a threshold.

IV. Result and Discussion

A. Clustering Results

The clustering experiment using AGRID+ algorithm has been done with the various numbers of cells (m) in the spatial dimensions; start from $m = 2$ until $m = 54$ with 2 as a bias. While in the temporal dimension is used the interval in 1 day (daily), 7 days (weekly), and 30 days (monthly). The cluster results are evaluated using silhouette index and analyzed what value of m that gives the better clustering result.

The graphic of silhouette index at different values of spatial m is shown in Fig. 2. From the clustering results, the best clustering obtained when using daily temporal with spatial m = 14 that give the silhouette index 0.9813. The utilization of optimal spatial m at three different temporal intervals is shown in Table I. Fig. 3 show the scatter diagram of the spatio-temporal fish catches data and its clustering result at daily temporal with m = 14.

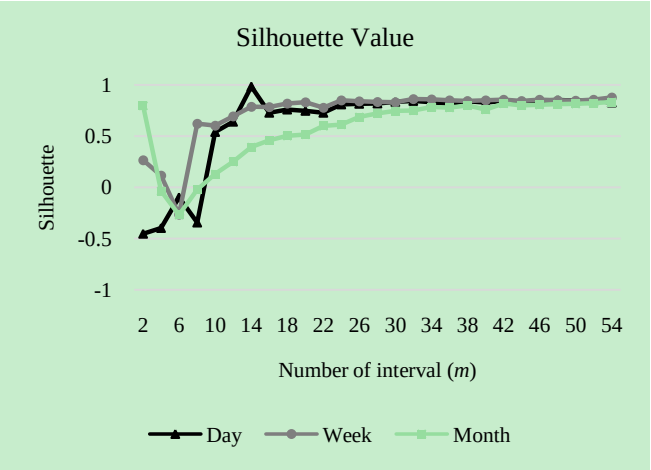
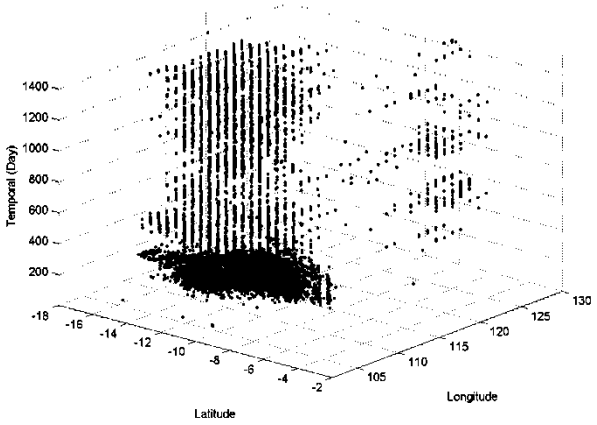
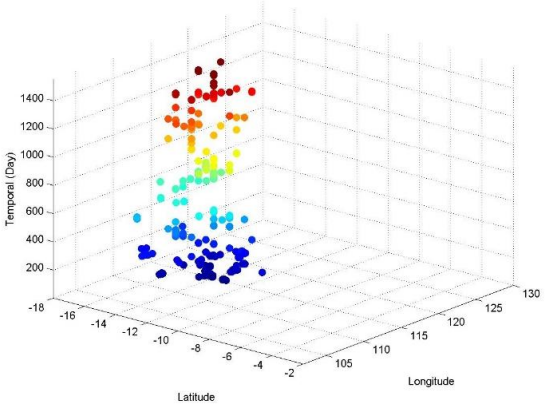


Fig. 2. The silhouette index of AGRID+ at three different temporal intervals.



a)



b)

Fig. 3. a) Data. b) Clustering result for daily temporal with m=14 and DT = 5.

TABLE I: The Silhouette Index of AGRID+

m temporal	*m* spatial	# of cluster	silhouette index
Daily	14	50	0.9813
7 days	54	676	0.8771
30 days	54	639	0.8317

TABLE II: The Average Execution Times

Temporal	Average *DT*	Average *r*	Exec.time (sec)	Silhoutte index
Daily	4.5311	0.0181	37.4082	0.6346
7 days	11.1777	0.0190	53.6410	0.7260
30 days	36.6741	0.0227	135.2700	0.5787

Fig. 4. The execution time of AGRID+.

In this experiment, we also analyze the execution time of clustering process with AGRID+ algorithm at three different temporal intervals. The graphic result is shown in Fig. 4, from that figure we can see that the constant execution time is in line with the increasing the value of *m*. From three different temporal aggregate, the daily aggregate is running relatively constant for various *m* values. The average execution times and silhouette values of the clustering results are shown in Table II.

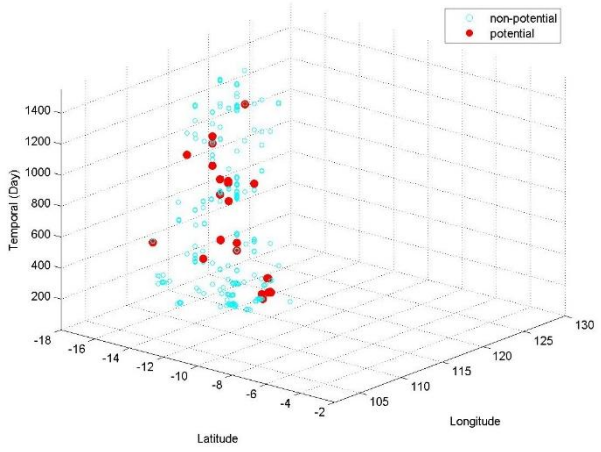
B. Potential Fishing Zone

Determination of the area that stated as PFZ can be done through thresholding technique. This threshold is obtained from the opiniion of a fishing company, PT. Perikanan Nusantara Indonesia, which states that an area categorized to be potential if the amount of the catch is equal to or grater than 5 in a single fishing trip. This number is related to the profits calculation with the minimum number of catches.

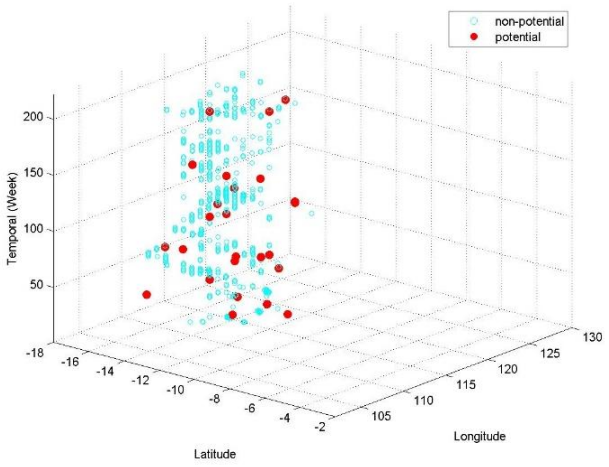
Referring to Equation (5), Table III and Fig. 5 shows the results and visualizations of PFZ with the amount of catches in that cluster areas are equal to or greater than 5 at three differents temporal intervals.

TABLE III: POTENTIAL FISHING ZONE ANALYSIS

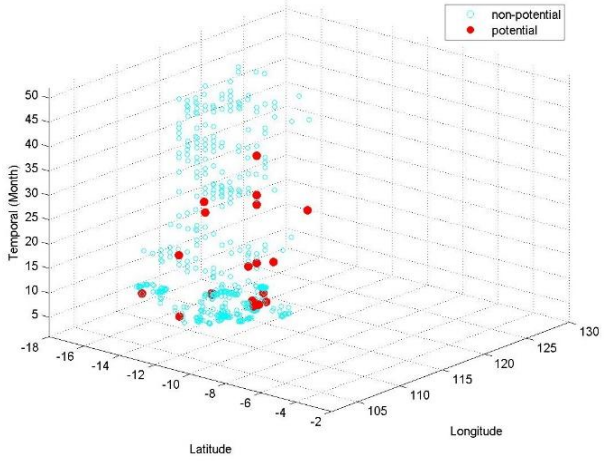
m temporal	*m* spatial	# of potential cluster	# of potential object
Daily	14	22	33
7 days	54	31	35
30 days	54	21	44



a)



b)



c)

Fig. 5. Potential fishing zone, a) daily, b) weekly, c) monthly.

V. CONCLUSION AND FUTURE WORKS

Potential fishing zone can be determined from spatio-temporal fish catches data using data mining approach. AGRID+, a grid-density based clustering, can be used to perform clustering data. The clustering results are used as a reference for determining PFZ using thresholding techniques.

The future works are to do an adaptation of AGRID+ algorithm especially for spatio-temporal data only. We also planned to do a fish forecasting by adding features of sea surface temperature (SST) and Chlorophyll-a.

REFERENCES

[1] J. A. Trinanes, J. M. Cotos, A. Tobar, and J. Arias, "A geographic information system for operational use in pelagic fisheries-FIS," in *Proc. OCEANS "94. 'Oceans Engineering for Today's Technology and Tomorrow" s Preservation*, 1994, pp. III/532–III/535.

[2] M. Sadly, N. Hendiarti, S. Sachoemar, and Y. Faisal, "Fishing ground prediction using a knowledge-based expert system geographical information system model in the South and Central Sulawesi coastal waters of Indonesia," *International Journal of Remote Sensing*, vol. 30, no. 24, pp. 6429–6440, 2009.

[3] M. Zainuddin, K. Saitoh, and M. Saitoh, "Albacore (Thunnus alalunga) fishing ground in relation to oceanographic conditions in western North Pacific Ocean using remotely sensed satellite data," *J. Fish. Oceanogr*, vol. 12, no. 2, 2006.

[4] S. Jagannathan, A. Samraj, and M. Rajavel, "Potential fishing zone estimation by rough cluster prediction," presented at 4th International Conference on Computational Intelligence, Modelling and Simulation, 2012.

[5] Y. Zhao, J. Cao, C. Zhang, and S. Zhang, "Enhancing grid-density based clustering for high dimensional data," *The Journal of Systems and Software*, vol. 84, pp. 1524-1539, 2011.

[6] S. K. Sahu and K. V. Mardia, "Recent trends in modeling spatio-temporal data," presented at Workshop on Recent Advances in Modeling Spatio-Temporal Data, 2005.
[7] T. Abraham and J. F. Roddick, "Opportunities for knowledge discovery in spatio-temporal information systems," *AJIS*, vol. 5, no. 2, 1998.
[8] Y. Zhao and J. Song, "AGRID: An efficient algorithm for clustering large high-dimensional datasets," in *Proc. the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 271-282, 2003.
[9] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

Devi Fitrianah was born in Jakarta, Indonesia in 1978. She received her bachelor degree in computer science from Bina Nusantara University, Indonesia in 2000 and master degree in information technology from Universitas Indonesia in 2008. She is currently pursuing the Ph.D degree in computer science at Faculty of Computer Science Universitas Indonesia.

At present, she is a research assistant in the image processing, pattern recognition and GIS laboratory. Her research interests are in image procesing, data mining, applied remote sensing and geographic information system.

Hisyam Fahmi was born in Malang, Indonesia in July 1989. He gets his bachelor degree in informatics engineering from Intitut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2011and Master degree in computer science from Universitas Indonesia in 2013.

Mr. Fahmi now is a research assistant in image Processing, pattern recognition and GIS laboratory, Faculty of Computer Science, Universitas Indonesia. His research interest is in image processing, pattern recognition, and machine learning.

Achmad Nizar Hidayanto was born in 1976. He received the B.S. degree in computer science from Universitas Indonesia, in 1999 and the M.S. degree in computer science from Universitas Indonesia, in 2002. He received his Ph.D. in computer science from Universitas Indonesia in 2008.

Dr. Hidayanto at present is appointed as the head of Information Systems Department, Faculty of Computer Science, Universitas Indonesia. He is an author of more than 40 articles written in journals and conferences. His research interests are related to information systems information technology, e-learning, information systems security, change management, distributed systems andinformation retrieval.

Aniati Murni Arymurthy is a professor at the Faculty of Computer Science, Universitas Indonesia. She was graduated from the Department of Electrical Engineering, Universitas Indonesia, Jakarta, Indonesia. She earned her master of science from the Department of Computer and Information Sciences, The Ohio State University (OSU), Columbus, Ohio, USA. She also holds doktor from Universitas Indonesia and a sandwich program at the Laboratory for Pattern Recognition and Image Processing (PRIP Lab), Department of Computer Science, Michigan State University (MSU), East Lansing, Michigan, USA. She is the head of Laboratory for Pattern Recognition and Image Processing, Faculty of Computer Science, Universitas Indonesia. Her research interests include the use of pattern recognition and image processing methods in several applications such as remote sensing, biomedical application, cultural artefak, agriculture and e-livestock.